Bayesian networks as a tool for epidemiological systems analysis

Lewis, F I

Abstract: Bayesian network analysis is a form of probabilistic modeling which derives from empirical data a directed acyclic graph (DAG) describing the dependency structure between random variables. Bayesian networks are increasingly finding application in areas such as computational and systems biology, and more recently in epidemiological analyses. The key distinction between standard empirical modeling approaches, such as generalised linear modeling, and Bayesian network analyses is that the latter attempts not only to identify statistically associated variables, but to additionally, and empirically, separate these into those directly and indirectly dependent with one or more outcome variables. Such discrimination is vastly more ambitious but has the potential to reveal far more about key features of complex disease systems. Applying Bayesian network modeling to biological and medical data has considerable computational demands, combined with the need to ensure robust model selection given the vast model space of possible DAGs. These challenges require the use of approximation techniques, such as the Laplace approximation, Markov chain Monte Carlo simulation and parametric bootstrapping, along with computational parallelization. A case study in structure discovery - identification of an optimal DAG for given data - is presented which uses additive Bayesian networks to explore veterinary disease data of industrial and medical relevance.

DOI: https://doi.org/10.1063/1.4765550

Bayesian Networks as a tool for Epidemiological Systems Analysis

F. I. Lewis

Section of Epidemiology, University of Zurich, Winterthurerstrasse 270, CH-8057 Zurich, Switzerland

Abstract. Bayesian network analysis is a form of probabilistic modeling which derives from empirical data a directed acyclic graph (DAG) describing the dependency structure between random variables. Bayesian networks are increasingly finding application in areas such as computational and systems biology, and more recently in epidemiological analyses. The key distinction between standard empirical modeling approaches, such as generalised linear modeling, and Bayesian network analyses is that the latter attempts not only to identify statistically associated variables, but to additionally, and empirically, separate these into those directly and indirectly dependent with one or more outcome variables. Such discrimination is vastly more ambitious but has the potential to reveal far more about key features of complex disease systems. Applying Bayesian network modeling to biological and medical data has considerable computational demands, combined with the need to ensure robust model selection given the vast model space of possible DAGs. These challenges require the use of approximation techniques, such as the Laplace approximation, Markov chain Monte Carlo simulation and parametric bootstrapping, along with computational parallelization. A case study in structure discovery - identification of an optimal DAG for given data - is presented which uses additive Bayesian networks to explore veterinary disease data of industrial and medical relevance.

Keywords: Bayesian networks; statistical modeling; machine learning; epidemiology
PACS: 87.10.Rt; 87.10.Mn

INTRODUCTION

Analysing observational data in order to provide insight into potential causes of disease, and factors associated with increased risk of exposure is extremely common in both human and veterinary medicine. From a data analysis perspective this is often far from trivial. Diseases and health conditions which are a priority for control or eradication in humans and animals are increasingly recognised to have highly complex determinants. For example, many diseases with high global burdens are endemic as a result of a multitude of different factors, some biological - properties of the pathogen - but also numerous inter-related social and economic conditions which provide an environment in which exposure is probable. Analysing observational data - realisations from the unknown stochastic processes which describe such epidemiological systems - is best served by a methodology which focuses neither on a single dependent variable, e.g. as in generalised linear modeling (GLM) with disease presence as the response variable, or some opaque dimension reduction technique such as factor or principal component analysis. Consider instead, additive Bayesian Networks (ABN), a form of graphical modeling which generalises the usual GLM to multiple dependent variables and involves no dimension reduction. We present here additive Bayesian Networks (ABN) as a data analysis tool for epidemiological systems analysis.

Statistical modeling of observational data presents considerable challenges; unlike in controlled experiments it is not generally possible to disentangle the effect of any individual covariate from another. Yet, in order to provide meaningful epidemiological interpretation of such data, attempts must be made to identify what potential relationships might be present between one or more outcomes of interest and risk factors (covariates). This is a long standing problem of major practical importance in epidemiology. Multivariable regression - one dependent variable and multiple independent variables - is by far the most commonly used statistical approach (e.g. [1, 2]). We suggest here that such approaches are not optimal when viewing the presence of disease as part of a complex system. By generalising GLMs into fully multi-dimensional models - multiple dependent variables - there exists the potential for considerably greater insight to be gained into complex epidemiological systems. The Yule-Simpson paradox [3] - that an apparent relationship between variables may disappear or even be reversed when others are taken into account - provides conceptual justification. An empirical justification is simply that results from an ABN analysis will generally differ from an analogous GLM analysis, where the former is simply a generalisation of the latter. A key distinction between much of the existing Bayesian network (BN) literature and ABN modeling, is that the former has focused largely on conjugate DAG models, thus retaining both mathematical elegance and computational efficiency. In the important case of binary/multinomial data this leads to a somewhat unhelpful contingency table parameterisation. In epidemiological analyses ready interpretation

9th International Conference on Mathematical Problems in Engineering, Aerospace and Sciences
AIP Conf. Proc. 1493, 610-617 (2012); doi: 10.1063/1.4765550

of the model and its parameters - e.g. odds ratios - are of paramount importance. This is arguably one reason why BN modeling is still rare in epidemiology despite its obvious potential for analysing observational data in medical and biological studies.

In the following sections we first provide a brief overview of modeling observational data using GLMs, BNs and ABNs, and how these relate to each other. This is followed by a veterinary case study which demonstrates how ABN modeling can be applied to data, and is contrasted with the use of standard multivariable regression approaches. We conclude with a discussion of the current limitations and challenges of ABN modeling and possible future directions.

MODELS FOR OBSERVATIONAL STUDIES

Observational data is a rather general term and is used here to denote any form of study where the allocation of subjects into different treatment groups is beyond the control of the investigator. Indeed, there may not even a pre-defined treatment or control group. A popular example, particularly in developing countries, are household surveys. The key point here is that the particular pattern of covariates which are observed - e.g. responses to questions in a questionnaire - are determined by unknown stochastic processes and not the investigator. By analysing observations from this system it is hoped to be able to elucidate some of the gross features of these underlying processes - thereby identify covariates which may be able to influence the health of the population.

Multivariable regression into which a variable selection process is then employed - typically a stepwise search - is the standard approach used in epidemiological studies concerned with identifying disease risk factors. The use of automated variable selection/model comparison techniques - algorithmic approaches in the terminology of Breiman [4] - have historically been viewed negatively in both the statistical and epidemiological literature. In contrast, and perhaps not entirely surprisingly, such automation is strongly embraced by the computer science and machine learning communities. Bayesian Networks and ABNs are very much in the spirit of Breiman's call for the preferred use of algorithmic techniques in data analysis.

Bayesian Networks

Bayesian network modeling is long established in the machine learning literature [5, 6]. Relatively recent methodological developments in the area of structure discovery - identifying optimal models (DAGs) for given data - include Markov chain Monte Carlo order-based searching [7] and exact dynamic programming approaches [8]. The application of BN models in biomedical science has been rather slow but is increasingly finding application in areas such as systems biology [9, 10], in HIV and influenza research [11–14], and also analyzes of complex disease systems [15–17].

Bayesian Networks Model definition

Figure 1) shows a DAG model for a four dimensional

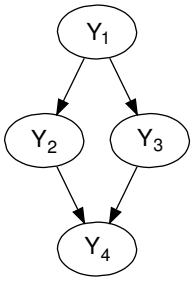

FIGURE 1. DAG model for four random variables, $Y_1,\ldots,Y_4$. A Bayesian network comprises of a DAG structure and a parameterisation. Let $Y_1,\ldots,Y_4$ denote binary variables and π_i for $i=1,\ldots,4$ denote the probability of observing a success: $P(Y_i=1)=1-P(Y_i=0)$. Y_1 is independent: $\log\{\pi_1/(1-\pi_1)\}=\beta_{1,0}$; Y_2 and Y_3 are conditionally dependent upon Y_1: $\log\{\pi_2/(1-\pi_2)\}=\beta_{2,0}+\beta_{2,1}Y_1$ and $\log\{\pi_3/(1-\pi_3)\}=\beta_{3,0}+\beta_{3,1}Y_1$. Y_4 is jointly dependent upon Y_2 and Y_3: $\log\{\pi_4/(1-\pi_4)\}=\beta_{4,0}+\beta_{4,1}Y_2+\beta_{4,2}Y_3$. This parameterisation gives an additive Bayesian network model.

joint probability distribution, $f(Y_1,Y_2,Y_3,Y_4)$, where the arcs denote conditional dependencies. In this example we have a factorization of $f(Y_1,Y_2,Y_3,Y_4)=f_1(Y_1)\ f_2(Y_2|Y_1)\ f_3(Y_3|Y_1)\ f_4(Y_4|Y_2,Y_3)$. Bayesian networks are decomposable [6] - each node can be considered separately - and we parametrize each node into a local logistic regression model where the goodness of fit and posterior parameters can be estimated independently within each node. The observed data for a logistic network model are tuples of the form $\{Y_1=d_1,\ldots,Y_m=d_m\}$, with $m=4$ in the example, where m is the total number of random variables (nodes) in the network model and $d_j\in\{0,1\}$ for $j=1,\ldots,m$ where $P(Y_j=1)=p_j=1-P(Y_j=0)$.

For ease of notation we use Y_j to denote the response variable at node j and those covariates on which Y_j is conditionally dependent (the nodes parents) as X_k, where these variables all belong to the same set $\{Y_1,\ldots,Y_m\}$. A logistic link function is used in each node where $p_{j,i}=\exp(\mathbf{x}^T_{\{j,i\}}\beta_j)/\{1+\exp(\mathbf{x}^T_{\{j,i\}}\beta_j)\}$ where $\mathbf{x}^T_{j,i}$ is the ith row in the design matrix $\mathbf{X}_j$ for node j, and $\mathbf{x}^T_{j,i}=[x_{j,i,0},\ldots,x_{j,i,m_j-1}]$ where $x_{j,i,k}$ denotes the value of the ith observation for the kth covariate in the logistic model

for node j and $x_{j,i,0} = 1 \ \forall i, j$ as a (separate) intercept term is included at each node. Again for simplicity of notation local indexing of the covariates at an individual node is used and is different from the global indexing of variables $j = 1, \ldots, m$. We use k for local indexing where k will always run from $k = 0, \ldots, m_j - 1$. The vector β_j denotes the coefficients for the model at node j and is of dimension m_j.

The likelihood function for node j is

$$L_j = \prod_{i=1}^{N} \left(\frac{\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)}{\{1+\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)\}} \right)^{y_{j,i}} \times \left(\frac{1}{\{1+\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)\}} \right)^{1-y_{j,i}}$$

where N is the total number of observations and complete data is assumed across all nodes (so N is constant $\forall j$). The total log-likelihood for the data is $l = \sum_{j=1}^{m} \log L_j$.

Priors

We develop our logistic network model within a Bayesian framework, and priors are therefore required for the parameters β_j for $j = 1, \ldots, m$, where these are the usual additive coefficients in a binomial GLM at each node. A number of different joint priors have been considered in conjugate categorical BN models, two commonly used metrics/priors are the Bayesian Dirichlet equivalence (BDe) metric and the K2 metric. It has been shown that the BDe is likelihood equivalent [6] which is theoretically desirable, in contrast, the K2 metric is not likelihood equivalent but does assume flat priors for all parameters, unlike with the BDe metric. Both metrics have been widely used in practice. To date we are unaware of any theoretical work in respect of developing likelihood equivalent parameter priors for non-conjugate BNs. In the absence of obvious alternatives and for simplicity we assume independent Gaussian priors for all parameters. In the subsequent case study analyses uninformative Gaussian priors with means of zero and variances of 1000 are used.

As we are searching across DAGs - to identify optimally fitting structures - there is also the need for a prior on structures. The default being that each structure is equally supported *a priori*. It is possible to construct informative structural priors, for example to penalize models with more structural complexity, e.g. more arcs, but as noted in [6] these are problematic to specify in practice. In [8] an informative structural prior on the number of parents within an individual node is used, where this assumes that parent combinations with the same cardinality are equally likely. This prior gives equal weighting to a parent combination with cardinality zero and cardinality $m-1$ which may not be entirely desirable. In the subsequent case study analyses an uninformative - flat - structural prior is used.

Goodness of fit measures

Structure discovery is all about model selection, and we consider the usual Bayesian goodness of fit metric, the marginal likelihood [18] - equivalent to Bayes factors for models with equal structural priors. In our additive logistic BN the marginal likelihood (conditional on a given DAG) for node j is

$$f_j(\mathbf{D}_j) = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{N} \left(\frac{\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)}{\{1+\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)\}} \right)^{y_{j,i}} \times \left(\frac{1}{\{1+\exp(\mathbf{x}_{\{j,i\}}^T \beta_j)\}} \right)^{1-y_{j,i}} \right] \times \prod_{k=0}^{m_j-1} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\{-(\beta_k - \mu_k)^2/(2\sigma_k^2)\} d\beta_j$$

where $\mathbf{D}_j$ denotes the observed data at node j and comprises of tuples of $[y_j, \mathbf{x}_{j,i}^T]$ and recall that vector $d\beta_j$ is of dimension m_j. The log marginal likelihood for the complete model is therefore $f(\mathbf{D}) = \sum_{j=1}^{m} \log f_j(\mathbf{D}_j)$.

The marginal posterior density for an individual parameter $\beta_{j,k}$ in node j can be estimated by successively evaluating equation $f_j(\mathbf{D}_j)$ for a fixed $\beta_{j,k} = b$ across the domain $-\infty < b < \infty$ and dividing by the normalizing constant $f_j(\mathbf{D}_j)$.

VETERINARY CASE STUDY

We now demonstrate the application of ABN structure discovery to data from a complex disease system. We break the structure discovery process into three sequential steps: i) determine an optimal DAG structure; ii) assess over-fitting of the structure in i) and if necessary prune excess complexity; iii) adjust the model in ii) for within group correlation structure (if applicable). We then have our "optimal" ABN model of the disease system. Ideally, step iii) would be combined into i) and we return to this complication later. In the following sections we illustrate how steps i) through iii) may be implemented. First we introduce the case study data then provide a simple comparison between using a GLM - equivalent to multivariable logistic regression - to explore this data and the more general ABN - equivalent to multivariate logistic regression.

Data

We utilize data on disease occurrence in pigs provided by the industry body the "British Pig Health Scheme" (BPHS). The main objective of BPHS is to improve the productivity of pig production in the UK, and reducing disease occurrence is a significant part of this process. The data we consider here comprise of a randomly chosen batch of 50 pigs from each of 500 randomly chosen pig producers in the UK. These are "finishing pigs", animals about to enter the human food chain at an abattoir. Each animal is assessed for the presence of a range of different disease conditions by a specialist swine veterinarian. We consider here the following nine disease conditions: enzootic-pneumonia (EPcat); pleurisy (plbinary); milk spots (MS); hepatic scarring (HS); pericarditis (PC); peritonitis (PT); lung abscess (Abscess); tail damage (TAIL); and papular dermatitis (PDcat). The presence of any of these conditions results in an economic loss to the producer. Either directly due to the relevant infected part of the animal being removed from the food chain, or indirectly in cases such as enzootic-pneumonia, which may potentially indicate poor herd health and efficiency losses on the farm. An additional loss, though not directly monetary, is the presence of tail damage which may be suggestive of welfare concerns, which may also be linked to sub-optimal production efficiency. Milk spots and hepatic scarring result from infestation with *Ascaris suum* which is particularly important as this is a zoonotic helminth parasite [19].

GLM or ABN

Figure 2 shows a globally optimal ABN for the case study data using the exact order-based search algorithm of [8] implemented in R [20] through the author's abn library (which is available for download from the CRAN website). Note here that there is no arc connecting PC and Abscess. The goodness of fit - log marginal likelihood - for this model is -44245.73. Forcing into this model an arc connecting PC and Abscess gives a far poorer log marginal likelihood of -44249.58 or -44249.51 (depending on the arc direction). Clearly this arc is not supported in this globally optimal DAG of the data.

Now consider a GLM analyses of the same data, where we consider PC and then Abscess as the response variables. The parameter and structural priors are identical to the ABN and a similar exact order-based search is conducted to determine a globally optimal structure. Figure 3 shows the two corresponding DAGs - a GLM is simply a DAG where arcs are only allowed directly between the covariates and response variable. In each case we find

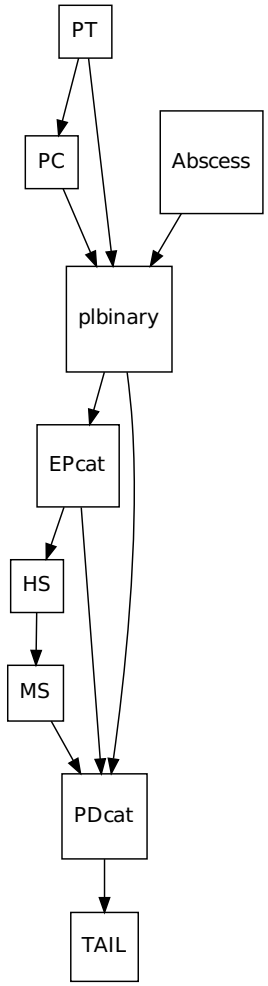

FIGURE 2. ABN model - optimal DAG - multivariate logistic regression of disease data

that an arc is identified between PC and Abscess. Moreover, the marginal posterior density in the GLM (panel a in Figure 3) for the arc from Abscess to PC strongly suggests that the log odds ratio here is highly significantly different from 0 (Figure 4 with 0 outside the 99.9 percentile). Using the reverse arc (panel b in Figure 3) gives virtually identical results.

In summary, we find that the GLM analyses identifies a strongly supported statistical association between presence of PC and the presence of Abscess. Applying an ABN model - a multivariate analogue of the GLM - to the same data we find again that an association between PC and Abscess is also supported but with a very important distinction. The ABN model does not support a direct

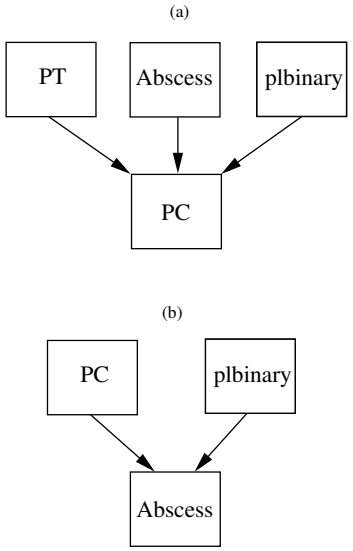

FIGURE 3. Two globally optimal GLMs - one with PC as the dependent variable (a), and a second with Abscess as the dependent variable (b).

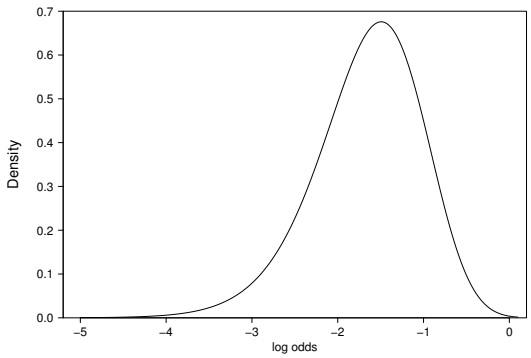

FIGURE 4. Marginal Posterior density for covariate Abscess and response variable PC

statistical dependency between PC and Abscess - there is no arc connecting these variables. Rather, the relationship (association) between PC and Abscess is via the intermediate variable plbinary. This highlights the key difference between a multivariable GLM and a multivariate GLM (ABN). The former identifies variables which may be associated with the response (dependent) variable within a very restrictive model space - arcs are only allowed from covariates direct to the response variable. When considering the same data within a larger model space, which incorporates other relationships within the underlying epidemiological system which generated the observed data, then such variables may then only be supported as indirectly - rather than directly - related to the response variable.

There is a profound difference in epidemiological interpretation between variables that are merely associated - indirectly dependent - from those which are directly dependent. The latter are natural targets for intervention strategies, the former - and whilst exceptions do exist - are typically of much lesser interest. This difference could be described as addressing lung cancer through targeting alcohol consumption, on the assumption that higher alcohol consumption is dependent with smoking, and smoking dependent with lung cancer. Alcohol consumption, therefore, is indirectly dependent with the presence of lung cancer. Obviously, such a strategy makes little sense compared to targeting smoking itself - as this would require a reduction in alcohol consumption to affect a reduction in smoking which then decreases the risk of lung cancer.

Ultimately, what is desired in epidemiological analyses is to identify variables which are on the casual pathway of the disease of interest. Whilst this is not possible using statistical analyses alone - as external information, such as a biological mechanism is required - intuitively, variables which are only indirectly dependent are less likely to be on such a pathway. Separating indirectly from directly dependent variables is, therefore, of considerable practical importance, and a strong justification for preferring an ABN over standard GLM type analyses.

In conclusion, the above example supports that an ABN approach is not only conceptually preferable for analysing data which arise from a complex disease system, but also that the resulting epidemiological interpretation of its results may also be substantively different from standard GLM approaches. As an ABN is simply a more flexible generalisation of a GLM then it is arguable that an ABN should always be used in preference, assuming it is computationally feasible to do so.

Structure Discovery

A particular challenge of ABN structure discovery is that it is NP-hard to find a globally optimal DAG [21]. This has led to the use of methods for iterating over orders rather than DAGs, however, problem size (number of variables) is still the main practical constraint in ABN analyses. In epidemiological applications the priority is typically to identify a single robust structure (e.g. [17]) as opposed to, for example, the use of model averaging approaches as clear and transparent interpretation of modeling results is essential.

i) Determine an optimal DAG. The globally optimal DAG for the case study data of 25000 pigs (Figure 2) was identified using an implementation of the order-based exact search algorithm of [8]. First, an exact search was conducted using a parent limit constraint of 1 - a

maximum of one arc to each node. A new exact search was run this time increasing the parent limit to 2, and the goodness of fit of these two models compared. If increasing the parent limit increased the goodness of fit, then the process was repeated until increasing the parent limit did not result in a model with improved fit. This iterative approach is usual to avoid the greatly increased computation time needed to search across the model space of DAGs with larger parent limits. The more direct approach of using a parent limit of $m-1$ could be used immediately but unless m is small then this search may be both extremely time consuming, indeed even practically infeasible, and is very inefficient unless a large enough amount of data is available to support such densely connected structures. For the case study data a parent limit of three was sufficient.

ii) Assess over-fitting. A parametric bootstrapping approach was suggested in [22] which uses simulation to assess whether a chosen model comprises more complexity than could reasonably be justified given the size of the observed data set. Using Markov chain Monte Carlo simulation via JAGS (open source software), 6000 independent (assumed by inspecting autocorrelations from the MCMC output) data sets of the same size as the original data were generated from our chosen model in i). For each of these bootstrap data sets an identical exact order-based search as in i) was conducted. Collating results across these 6000 searches we find that only 14% of the globally optimal DAGs found comprised 12 or more arcs. Approximately 68% of DAGs had 11 or more arcs - therefore a robust model of the original data has no more than 11 arcs. Almost identical results were obtained using a random selection of 3000 searches suggesting that sufficient bootstrap samples had been performed. The usual cut-off for structural support of features (arcs) is 50% in BN modeling (see [11–14, 16]), and is analogous to the widespread use of majority consensus trees in phylogenetics. We therefore conclude that our chosen model in i) with 11 arcs is robust. This is perhaps not surprising given we have a large data set of 25K observations.

iii) Adjustment for clustering/correlated observations. As in many veterinary studies we have a potential correlation structure within our 25K observations, as these are grouped observations from 500 different pig producers. This means that over-dispersion may be present in the data, as while two producers may have identical covariate patterns there are other aspects unique to their farms which may increase the level of variation beyond what is possible under binomial sampling. The usual way to resolve this is to move from a GLM to a GLMM, a generalised linear mixed model - a GLM which has random effects now included - which allows for an increase in variance and within group correlations [23]. The practical impact of such clustering being that some of the arcs previously identified in steps i) and ii) may no longer be supported after any additional variance due to within group correlations is included. We use MCMC to fit the model from ii) to the observed data where now an independent random effect is included at each node in the model, where these are all assumed to be Gaussian distributed with means of zero and diffuse Gamma distributed precision parameters (shape=0.001 and scale=0.001).

To determine whether any arcs should be dropped, marginal posterior 95% confidence intervals of the log odds ratio for each parameter were estimated. If this interval included zero then it was deemed that the corresponding arc did not have sufficient statistical support to be retained in the model. This resulted in four arcs being removed from the optimal model identified in step ii) and gives us a final ABN model (Figure 5).

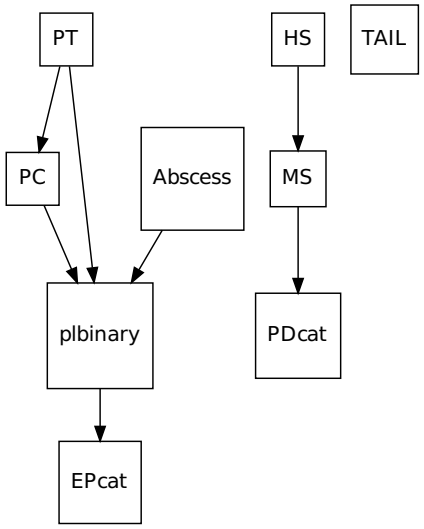

FIGURE 5. Final ABN model of disease data after bootstrapping and clustering adjustment

Our final "optimal" model of this disease system suggests that the presence of tail biting is independent from the other conditions. We also find that the remaining diseases are split into two separate connected components. This raises the interesting biological question as to whether these two groups of diseases may lie on different causal pathways, i.e. might the diseases within each of these groups share common causes. Such results from data analyses using ABN modeling can be used as a basis for developing new biological questions about factors potentially affecting the presence of disease, and inform the design of future targeted studies.

DISCUSSION

Two challenges of ABN structure discovery are computational robustness and computationally feasibility. Structure discovery requires automated and reliable numerical approximation applied to vast numbers of different models. Currently, a Laplace approximation [24] is used to compute the marginal likelihood for each and every DAG. To date, we have not included random effects in this estimation process, but rather have resorted to the somewhat easier, but less satisfactory situation, of requiring a third step in our model search process. This instead uses MCMC simulation, which while easy to implement, can take a very long time indeed to produce reliable output when dealing with numerous variables and numerous random effects. It also does not readily offer accurate computation of marginal likelihood (although see [25] for one option). A far better approach would be to include ABNs with random effects into step i) in the original model search process. This work is ongoing and it is as yet an open question as to how challenging the inclusion of multiple random effects will be in producing a robust automated algorithm for computing the marginal likelihoods across many different models comprising of numerous random effect terms. An easier alternative might be to use compound distributions such as the beta-binomial rather than explicit inclusion of random effects.

In terms of computational feasibility, the current implementation of Koivisto's [8] dynamic programming algorithm is readily feasible up to 20 or so nodes (variables). Beyond this it becomes rapidly challenging even on multi-core (cluster computer) hardware. The abn library in R includes some simple multi-threaded algorithms (using openMP) which provides a considerable speed-up in computation, but beyond 25 nodes this again becomes infeasible. It has been demonstrated that it is computationally possible to perform structural searches on up to approximately 40 nodes with specialist cluster specific algorithms. However, this is not a practical solution for real world epidemiological applications; if performing an exact search once takes a vast amount of computing then nesting this inside parametric bootstrapping is clearly not going to be feasible. This is an important constraint as, generally speaking, BNs tend to considerably over-fit data unless dealing with very large sample sizes.

In conclusion, data analyses using additive Bayesian networks have the potential to offer new insights into complex epidemiological systems. While there are many exciting computational challenges yet to be addressed this approach is still feasible for many veterinary and human medical studies.

ACKNOWLEDGMENTS

Manuel Sanchez-Vazquez provided invaluable assistance with the data and expert veterinary input. The British Pig Executive (BPEX) kindly allowed use of their data. Access to the Schrödinger scientific computing resource was provided by the University of Zürich.

REFERENCES

1. G. Phillips, B. Lopman, L. Rodrigues, and C. Tam, *American Journal of Epidemiology* **171**, 1023–1030 (2010), URL `://000276999100009`.
2. H. L. Johnson, L. Liu, C. Fischer-Walker, and R. E. Black, *International Journal of Epidemiology* **39**, 1103–1114 (2010).
3. D. J. Hand, K. J. McConway, and E. Stanghellini, *IMA Journal of Management Mathematics* **8**, 143–155 (1997), URL `http://imaman.oxfordjournals.org/content/8/2/143.abstract`, `http://imaman.oxfordjournals.org/content/8/2/143.full.pdf+html`.
4. L. Breiman, *Statistical Science* **16**, 199–215 (2001), ISSN 08834237, URL `http://www.jstor.org/stable/2676681`.
5. W. Buntine, "Theory refinement on Bayesian networks," in *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, Los Angeles, CA, USA., 1991, pp. 52–60.
6. D. Heckerman, D. Geiger, and D. M. Chickering, *Machine Learning* **20**, 197–243 (1995).
7. N. Friedman, and D. Koller, *Machine Learning* **50**, 95–125 (2003).
8. M. Koivisto, and K. Sood, *Journal of Machine Learning Research* **5**, 549–573 (2004).
9. C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead, *PLoS Comput Biol* **3**, e129 (2007).
10. A. Djebbari, and J. Quackenbush, *Bmc Systems Biology* **2**, 57 (2008).
11. A. F. Y. Poon, F. I. Lewis, S. L. K. Pond, and S. D. W. Frost, *Plos Computational Biology* **3**, 110–119 (2007).
12. A. F. Y. Poon, F. I. Lewis, S. L. K. Pond, and S. D. W. Frost, *Plos Computational Biology* **3**, 2279–2290 (2007).
13. A. F. Y. Poon, F. I. Lewis, S. D. W. Frost, and S. L. K. Pond, *Bioinformatics* **24**, 1949–1950 (2008).
14. S. J. Lycett, M. J. Ward, F. I. Lewis, A. F. Y. Poon, S. L. K. Pond, and A. J. L. Brown, *Journal of Virology* **83**, 9901–9910 (2009).
15. N. Dojer, A. Gambin, A. Mizera, B. Wilczynski, and J. Tiuryn, *Bmc Bioinformatics* **7**, 249 (2006).
16. F. I. Lewis, F. Brulisauer, and G. J. Gunn, *Preventive Veterinary Medicine* **100**, 109–115 (2011).
17. F. I. Lewis, and B. J. J. McCormick, *American Journal of E* **http://dx.doi.org/10.1093/aje/KWS183** (2012).
18. D. J. C. Mackay, *Neural Computation* **4**, 415–447 (1992).
19. C. Dold, and C. V. Holland, *Microbes and Infection* **13**, 632–637 (2011).
20. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2006), URL

http://www.R-project.org, ISBN 3-900051-07-0.
21. D. M. Chickering, D. Heckerman, and C. Meek, *Journal Of Machine Learning Research* **5**, 1287–1330 (2004).
22. N. Friedman, M. Goldszmidt, and A. Wyner, "Data analysis with Bayesian networks: A bootstrap approach.," in *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99) (pp.206-215). San Francisco: Morgan Kaufmann*, 1999.
23. C. E. McCulloch, S. R. Searle, and J. M. Neuhaus, *Generalized, Linear, and Mixed Models*, Wiley, 2008.
24. L. Tierney, and J. B. Kadane, *Journal of the American Statistical Association* **81**, 82–86 (1986).
25. S. M. Lewis, and A. E. Raftery, *Journal Of The American Statistical Association* **92**, 648–655 (1997).